Adversarial Robustness Gains in Rationale-Augmented DPO via Cross-Domain Fine-Tuning

Assignee Research

June 12, 2026

Abstract

State-of-the-art few-shot learning (FSL) methods leverage prompt-based fine-tuning to obtain remarkable results for natural language understanding (NLU) tasks. While much of the prior FSL methods focus on improving downstream task performance, there is a limited understanding of the adversarial robustness of such methods. In this work, we conduct an extensive study of several state-of-the-art FSL methods to assess their robustness to adversarial perturbations. To better understand the impact of various factors towards robustness (or the lack of it), we evaluate prompt-based FSL methods against

1 Introduction

This paper examines: Adversarial Robustness of Prompt-based Few-Shot Learning for Natural Language Understanding. Research question: What is the impact of cross-domain fine-tuning on the adversarial robustness gains of rationale-augmented DPO, evaluated on AdvBench and other robustness benchmarks like RobustBench?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.0/10.

3 Results

16 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 9.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Using unlabeled data (iPET) during fine-tuning causes prompting to reduce the drop in adversarial performance with respe	✓	0.27
Using multiple prompts to fine-tune multiple models (PET) and ensembling the resultant predictions cause prompting to de	✓	0.40
Increasing the number of few-shot examples and the encoder size reduces the relative drop in adversarial performance wit	✓	0.31
RoBERTa encoders are more adversarially robust than ALBERT and BERT encoders of comparable size.	✓	0.18
Few-shot learning aims to train models to perform well on a wide range of natural language understanding tasks with a sm	✓	0.29
Prompt-based learning overcomes the requirement of training task-specific classification heads, matching the fine-tuning	✓	0.24
FewNLU is a benchmark designed to evaluate the performance of prompt-based few-shot learning capabilities systematically	✓	0.22
Vanilla FSL methods lead to a notable relative drop in task performance (i.e., are less robust) in the face of adversari	✓	0.43
Using unlabeled data for prompt-based FSL and multiple prompts flip the trend of reduced robustness in the face of adver	✓	0.28
Increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL method	✓	0.42

References

- http://arxiv.org/abs/1905.11736v5
- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/2306.11066v2